Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200190

Prototyping a Tool for Processing Genetic Meta-Data in Microbiological Laboratories

Jan AESCHIMANN[a], Silvan HUBER[a], Daniel WÜTHRICH[b,c], Helena SETH-SMITH[b,c], Jürgen HOLM[a], Thomas BÜRKLE[a], and Murat SARIYAR[a,1]
[a] *Bern University of Appl. Sciences, Dept. Medical Informatics, Switzerland*
[b] *Clinical Bacteriology and Mycology, University Hospital Basel, Switzerland*
[c] *Applied Microbiology Research, Department of Biomedicine, University of Basel, Switzerland*

Abstract. Next generation sequencing (NGS) technologies allow improved understanding of pathogens. In the upstream processing of generating genomic data, there is still a lack of process-oriented tools for managing corresponding meta data. In this paper, we provide a description of how a process-oriented software prototype was developed that allowed the capture and collation of metadata involved when doing NGS. Our question was: How to develop an interactive web application that supports the process-oriented management of genetic data independent of any sequencing technique?

Keywords. data management; microbiological data; ngs; web application

1. Introduction

Modern microbiological laboratories use next generation sequencing (NGS) for understanding the genomes of pathogens. This allows, for example, the detection of outbreaks and the determination of antibiotic resistance mechanisms. In order to generate annotated genomic data, bioinformatic and biomedical knowledge are necessary. For NGS analysis, software solutions such as Ridom Seqsphere+ [1], CLC Genomics Workbench [2] or UGENE [3] can be used and facilitate the analysis of sequence data, annotation, alignments, and visualization such as phylogenetic trees. However, there is more information involved in the upstream processing of generating genomic data, which are often not fully considered in such approaches [4].

The collection and reliable storage of metadata relating to the sample is critical, particularly in an ISO accredited clinical microbiology laboratory, but also of relevance to other laboratories performing NGS. Three steps for generating a genome-based laboratory typing report from a clinical sample can be distinguished: sample preparation, sequencing and analysis. In the sample preparation step, clinical samples or isolates are subject to DNA extraction. In this stage, relevant data relating to the patient, the sample, and the extraction are collected. In the sequencing step, data on the concentration of DNA, the library method, the sequencing platform used, and a quality

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

control of the resulting sequences are annotated. In the analysis step, the sequences are interpreted, for which biomedical information is indispensable.

The workflow of capturing and collating metadata during NGS was analyzed in the Division of Bacteriology at the University Hospital Basel. The processes for capturing the data associated with NGS were examined before starting the process of software development. At baseline, Excel worksheets on a server were used with no established workflow of how and when to commit changes or new entries. Our central goals were to facilitate NGS metadata management in order to reduce the likelihood of errors in the database and allow for parallel editing of different parts of the database. The guiding research questions was: How to develop an interactive web application that allows an efficient process-oriented management of genetic data independent of any sequencing technique?

2. Methods

With a Pubmed and Google Scholar research using the keywords “metadata microbiology”, “ngs metadata”, “pathogen ngs metadata”, “metadata management system” and “metadata management system genome” we checked if an external tool for our goals is available. Next, we iteratively created a mockup together with the laboratory staff using the Figma mockup tool [5] in order to define an optimal workflow. The subsequent requirement engineering process was based on IEEE 830-1998 [6] and lead to the definition of the software architecture.

In the development phase, we used the agile SCRUM technique [7] with seven sprints lasting one week each. We made use of the JavaScript node.js (loopback) framework for the backend and the Vue.js framework for the frontend. The REST-based philosophy of many JavaScript frameworks is not designed for multiuser systems that require isolated updates and deletion of common resources. Therefore, we used *socket.io* as a JavaScript library for real-time web applications that enables bi-directional communication between web clients and servers, ensuring that locks on datasets can be propagated beyond one session within a REST-based access [8]. For usability tests based on a structured questionnaire and for final modifications according to the anwers, two more weeks were scheduled.

3. Results

In our analysis of the management of genetic meta data, we identified six central process steps concerning the weekly sample preparation workflow. Initially, a laboratory order from internal or external customers arrives together with the material to be sequenced. The order is validated and prioritized. In step 2, the sample type (isolate or DNA) is checked. Step 3 selects the necessary DNA extraction process with respect to the material, if required. Step 4 selects the number of samples (4, 24, 48, or 96 depending on the sequencing device and cartridge). After performing the DNA extraction, the sequencing can be planned with the information generated during extraction (step 5). Finally, the samples for the next NGS run are selected, a sequencing run is performed, and sequencing data is saved for later analysis and reporting (step 6).

These process steps are divided into four different views according to the data input required. For steps 1-2, it is the view "*Planned*" with data on the requesting facility, the patient, the type of the material, the date of receiving the order, the type of the pathogen, the priority of the order (from "A": high to "D": low); the view for step 3 is "*extracted*" with the same data as in the former view and DNA concentration; the view for steps 4 and 5 is "*Preparing run*" with data on the finalized DNA extracts to be sequenced; for the last step the view is "*Sequenced*" with data on the sequencing device used, date of sequencing and some of the sequencing properties, e.g. adaptor oligonucleotides ligated to fragments of the DNA for forming the DNA library.

During the requirements analysis, we also identified the original data management workflow based on Excel worksheets, which showed several disadvantages:

- Lack of information on the order of inserting data into the Excel worksheets
- No import options for pre- formulated orders, potentially leading to errors in manual entries
- No export options for selected data, which could lead to errors in manual copy/paste operations during report generation
- No support for batch processing of the data
- No possibility of representing a complete repetition of a sequence run by shifting metadata from the view "*sequenced*" to former views

To implement a web application that addresses these disadvantages and provides an affirmative answer to our research questions (see introduction), we considered additional functional and non-functional requirements, e.g. a table view for general overview or ubiquitous access to the system from everywhere within the laboratory, which mandates a web interface that is compatible with current browsers. The main non-functional requirement was user guidance through the metadata management process by prompting only for relevant fields in each step and by providing help functions, sorting options, search functions as well as tooltips. The prototype is called "Red Maple", and the start page is given in Figure 1. The UML sequence diagram for our solution using web sockets is presented in Figure 2.

Figure 1. Start page of the prototype *Red Maple* with the four views "Planned", "Extracted", "Run", "Sequenced" and a summary of datasets included in them. The bubbles in the right column represent the links for details within these views.

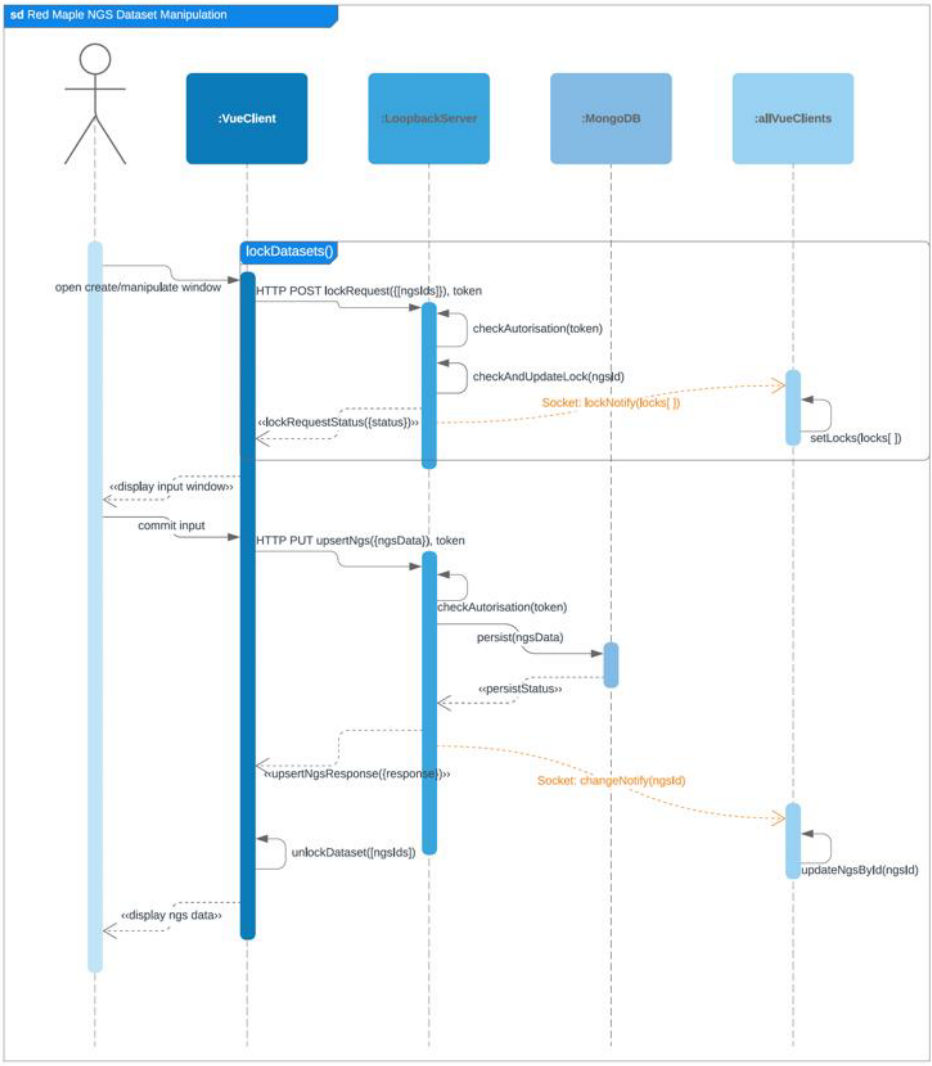

Figure 2. The UML sequence diagram shows how data is accessed via the frontend (VueClient). In case of manipulation of the data, tokens are used, which allow manipulation of data in the backend (MongoDB is the NoSQL database used by the loopback Server). Locks are necessary in order prevent anomalies possible when the same data is tried to be manipulated simultaneously.

4. Discussion

Typically, laboratory analyses are associated with clear diagnostic procedure steps described in medical guidelines. The NGS processes within microbiological laboratories are different in that they are often not directly based on a medical guideline but are necessary to gain information on transmission and outbreaks, thus affecting

more than one single patient. This is one of the reasons why such NGS analysis processes are difficult to integrate into the workflow of existing laboratory information systems. Hence, instead of stretching existing LIS solutions, it seems desirable to develop a modular solution that is useful for many microbiological laboratories. Therefore, we implemented and described an exemplary web application that supports an efficient process-oriented management of NGS-related meta data.

One important advantage of our REST-based prototype is the increasing number of RESTful applications and clients that can be easily interfaced in future, for example the biobank software *OpenSpecimen* [9]. HL7 FHIR is an important standard for RESTful medical software applications that makes RESTful service-oriented architectures increasingly attractive for the data management of a variety of "non-standard" clinical data.

Due to project time constraints, there are some limitations. Today, the prototype is not yet in routine use. We collected positive reactions from future users and measured decreased processing time, but no systematic study of the effects of the new solution has been conducted. Future work will also concentrate upon more sophisticated statistical reports for the data.

References

[1] Seth-Smith HMB et al. Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation. Front Public Health. 2019;27(7):241.
[2] Kim KU et al. Comparison of functional gene annotation of Toxascaris leonina and Toxocara canis using CLC genomics workbench. Korean J Parasitol. 2013;51(5):525-30
[3] Protsyuk IV et al. Shared bioinformatics databases within the Unipro UGENE platform. *J Integr Bioinform.* 2015;12(1):257.
[4] Hong EL et al. Principles of metadata organization at the ENCODE data coordination center. Database 2016. 2016;baw001.
[5] Figma Design. Figma: the collaborative interface design tool, Retrieved October 14, 2019 from https://www.figma.com/.
[6] IEEE Computer Society Software Engineering Standards Committee. IEEE Recommended Practice for Software Requirements Specifications, IEEE Standard 830-1998, 1998.
[7] Cervone HF. Understanding agile project management methods using scrum. OCLC Syst Serv. 2011;27(1):18-22.
[8] Hasibuan A et al. Design and implementation of modular home automation based on wireless network, REST API, and WebSocket, In: Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) - IEEE 2015. 2015; 362–367.
[9] McIntosh LD et al. caTissue Suite to OpenSpecimen: Developing an extensible, open source, web-based biobanking management system. J Biomed Inform. 2015; 57:456-464.